Anderson, Christopher

PREDICT 456 Section 55

Assignment #4

Introduction

The most elemental part of baseball at any level is the battle between pitcher and batter. This interaction sets in motion everything else that occurs in the game. Various pitch types are employed by pitchers as they attempt to induce swinging strikes and weakly hit groundouts or popups from batters. Common pitch types include the fastball, curveball, and changeup. In the past, pitches like the spitball and screwball have risen to prominence in Major League Baseball and subsequently faded. More recently, the slider and cut fastball, or cutter, have gained popularity. A glossary defining each pitch type is included in Appendix B. Since the pitcher's goal is always to get the batter out, it can be assumed that MLB pitchers aim to throw the types of pitches that are most effective in accomplishing that feat.

The effectiveness of different pitch types has been studied in the past, but the focus has primarily been on specific pitch types thrown by individual pitchers (e.g. Pitcher A's fastball or Pitcher B's curveball). Prominent baseball analyst and data journalist Eno Sarris studied the best pitches in baseball during the 2015 season and developed a scoring system based on swinging strike rate and ground ball rate, arguing that these are two of the most important outcomes a pitch can have (Sarris, 2016). He used it to rank pitches per player per pitch type, but noted that many other factors make a particular pitch effective. His method found that the best pitch during the 2015 season was a particular relief pitcher's sinker, but it was closely followed by other pitch types including curveballs, sliders, and fastballs. Furthermore, a report published by sports data and technology company STATS used batting average against to rank the best individual pitches in the game during the 2015 season (STATS, 2015). The top ten pitches produced by this method included mostly off-speed pitches such as sliders, curveballs, and changeups. Jeff Sullivan, a prominent writer at baseball statistics and analysis website Fangraphs.com, opted to use lowest contact rate to identify the most "unhittable" pitches in the league in 2013 (Sullivan, 2013). Similar to the results produced by STATS, the best pitches according to this approach were sliders, curveballs, and changeups.

Despite all of this research, Major league players and managers still assert that a well-placed fastball is the most difficult pitch to hit (Berry, 2013). Statistical expert Mike Fast, currently employed by MLB's Houston Astros as an Analyst, investigated the correlation between fastball velocity and performance, discovering that starting pitchers improve by one run allowed per nine innings for every 4 mph increase on fastball velocity (Fast, 2010). Relief pitchers achieve the same gain in performance for every 2.5 mph increase. It appears that one can determine that any particular pitch type is the best, depending on the metrics used in the assessment.

The purpose of this study is to examine the differences between pitch types in Major League Baseball and identify which pitch types have performed better than others during the 2016 season, as measured by batted ball hit speed, and opponent batting average on balls in play (BABIP). Its implications could impact areas such as scouting, roster construction, and player development. The data used in this report were generated by PITCHf/x, an advanced technology and software system that tracks nearly everything that occurs on the field during a game. Created in the early 2000s, it has been installed in every MLB stadium since the beginning of the 2008 season, tracking data such as the velocity and acceleration of the ball for each game (Marchi, 2013). Based on evidence presented above, the hypotheses for this study are: 1) slower pitch types such as curveball or changeup will perform better on batted ball hit speed; and 2) the downward breaking sinker will perform better on BABIP.

Methods

PITCHf/x data for every MLB game during the 2016 season through August 20th was extracted from the Statcast Search page on MLB's website. This resulted in a dataset of 541,581 observations across 60 variables. A subset of the data was taken to eliminate irrelevant variables and 13,095 incomplete records were removed, leaving 528,486 remaining observations. There was no explanation for the missing data upon downloading the data, but the most likely explanation is that the PITCHf/x technology failed to capture some aspect of the trajectory for those particular pitches – which accounted for only 2.4% of the dataset. The 12 variables used in the study are included in Table 1 below, along with a description and example of each.

Table 1: Variable descriptions

Variable	Description	Type	Example
pitch_type	Abbreviation indicating the type of pitch thrown	Nominal	FF
description	Description of the pitch type	Nominal	Ball
events	Description of the result of the at bat	Nominal	Single
type	An abbreviation indicating the result of the pitch	Nominal	B
start_speed	Velocity of the pitch measured at the initial point (MPH)	Ratio	96.10
effective_speed	"Perceived velocity" or how fast the pitch appears to a hitter (MPH)	Ratio	94.74
spin_rate	Revolutions per minute of the ball after it is released by the pitcher (MPH)	Ratio	1996.299
break_angle	Angle from the release point to where the pitch crossed the front of home plate	Ratio	-4.0
break_length	Greatest distance between the trajectory of the pitch and the straight line path from the release point and the front of home plate (inches)	Ratio	2.9
hit_speed	Speed of the ball after being struck by the bat (MPH)	Ratio	94.06
stand	Abbreviation indicating the side of the plate from which the batter hits	Nominal	R
p_throws	Abbreviation indicating the hand with which the pitcher throws	Nominal	L

Of particular importance is pitch type, which indicates the type of pitch thrown and serves as the independent variable throughout this study. Start speed is the traditional measurement of velocity used throughout baseball history. Effective speed represents how fast the pitch appears to the batter and is calculated based on the pitcher's release extension, or how far towards home plate their arm is when the pitch is thrown. Spin rate is an additional PITCHf/x field that represents the rate of spin on the ball after it is released by the pitcher (Glossary, 2016). Break angle and break length are measurements of the ball's trajectory that the PITCHf/x system uses to determine the pitch type (Fast, 2007). Hit speed measures the speed of the ball after it is struck by the bat (Glossary, 2016). For pitchers, a lower value of hit speed is desired. The mean values for each of these variables, per pitch type, are shown below in Table 2.

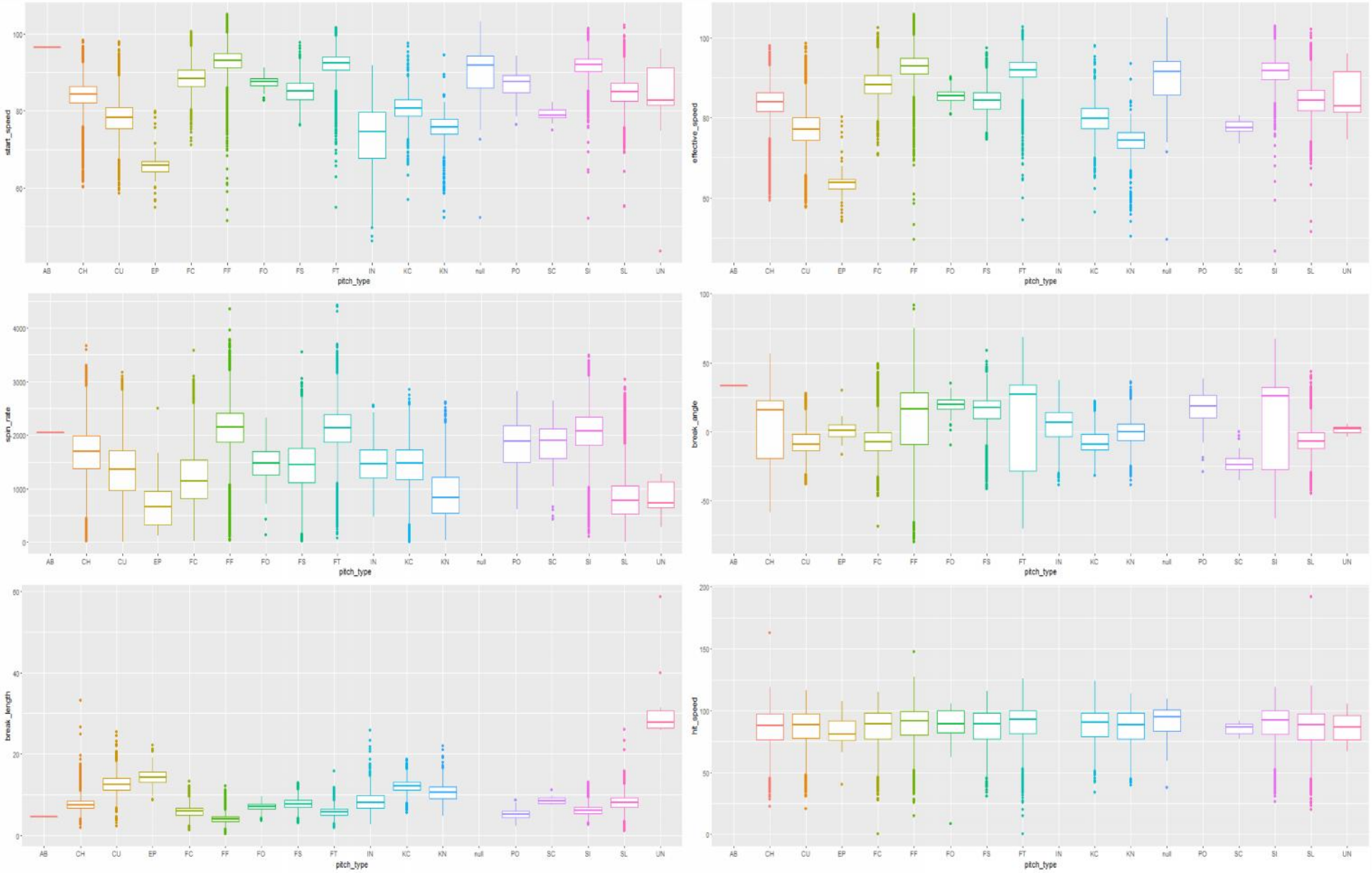
Table 2: Mean values by pitch type

pitch_type	count	start_speed	effective_speed	spin_rate	break_angle	break_length	hit_speed
CH	53901	83.94	83.69	1659.626	5.0	7.5	86.06
CU	44967	77.99	77.03	1339.254	-7.0	12.5	86.34
EP	81	65.21	63.10	665.189	-1.2	14.6	82.25
FC	26848	88.41	88.20	1170.314	-7.0	6.0	86.71
FF	190461	93.00	92.77	2085.628	8.2	4.2	89.21
FO	178	87.63	85.76	1334.953	17.6	7.0	86.35
FS	7708	84.95	84.15	1370.424	12.8	7.8	86.88
FT	72420	92.32	91.94	2094.365	11.0	5.8	89.79
KC	11642	80.69	79.73	1414.534	-7.3	12.1	87.76
KN	3730	75.75	74.28	846.728	-0.3	10.7	86.18
SC	25	79.60	78.29	1877.050	-24.1	8.3	85.08
SI	34676	91.72	91.51	2042.070	11.4	6.3	89.53
SL	80589	84.96	84.39	822.462	-6.2	8.1	86.27
UN	51	89.86	88.84	593.752	1.4	29.1	86.32

Some pitch types are simple and thrown by nearly every pitcher while others are difficult to master and rarely put to use. It is clear that the four-seam fastball (FF) was by far the most commonly thrown pitch during the period, with other variations of the fastball also among the most heavily used. On the other hand, the screwball (SC) was the most seldom used pitch type with only 25 occurrences in 2016. Also, it was evident from the large variation in speed, spin rate, break angle, and break length that the characteristics of each pitch type are rather unique.

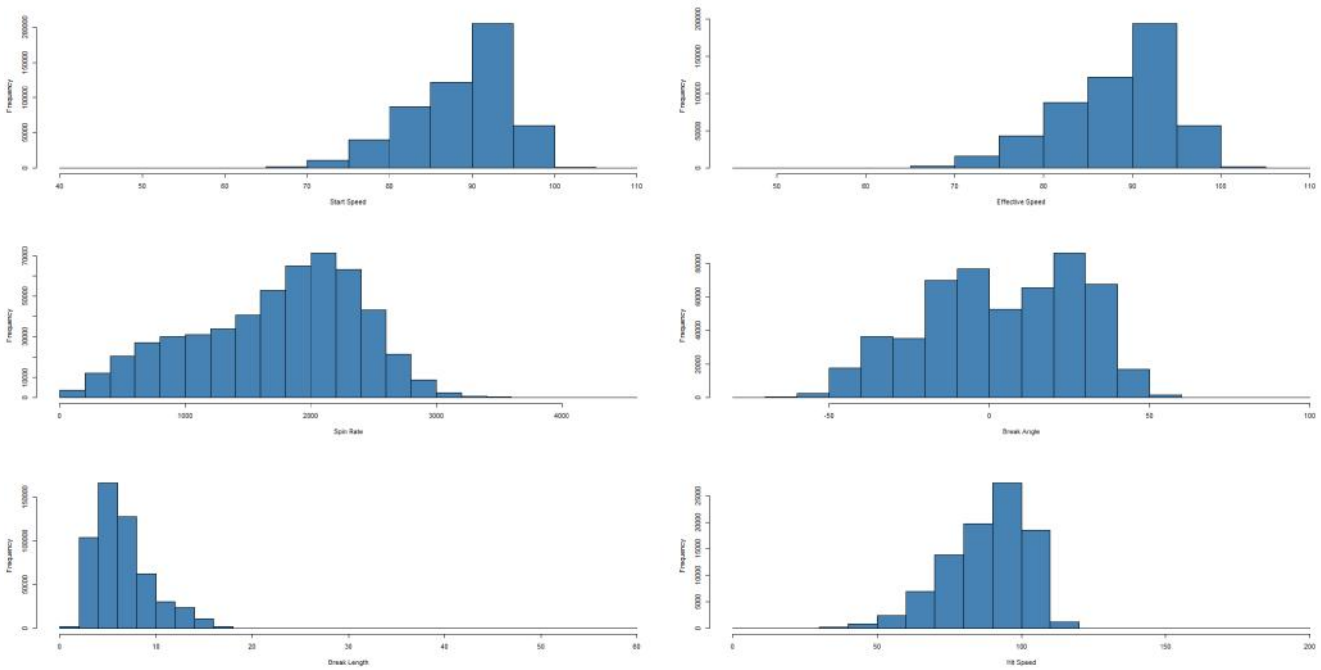
Figure 1 on the following page presents boxplots of each variable by pitch type. These plots illustrated how the trajectories vary drastically for each pitch type. The disparities exhibited below suggested that the differences between pitch types may be significant. Conversely, the boxplot for hit speed appeared to be fairly uniform across all pitch types. However, slight variations in hit speed can make a substantial difference in baseball so a statistical test was used to determine if there were, in fact, differences between groups for hit speed.

Figure 1: Boxplots by pitch type



An analysis of variance (ANOVA) is commonly used when there are one or more categorical independent variables and one continuous dependent variable. A key assumption for the ANOVA test is that the dependent variable is normally distributed within each group that is being compared. The histograms presented below in Figure 2 illustrate that the distributions for start speed, effective speed, spin rate, break angle, break length, and hit speed are decidedly not normally distributed. There are indications of both positive and negative skewness, as well as bimodal distributions.

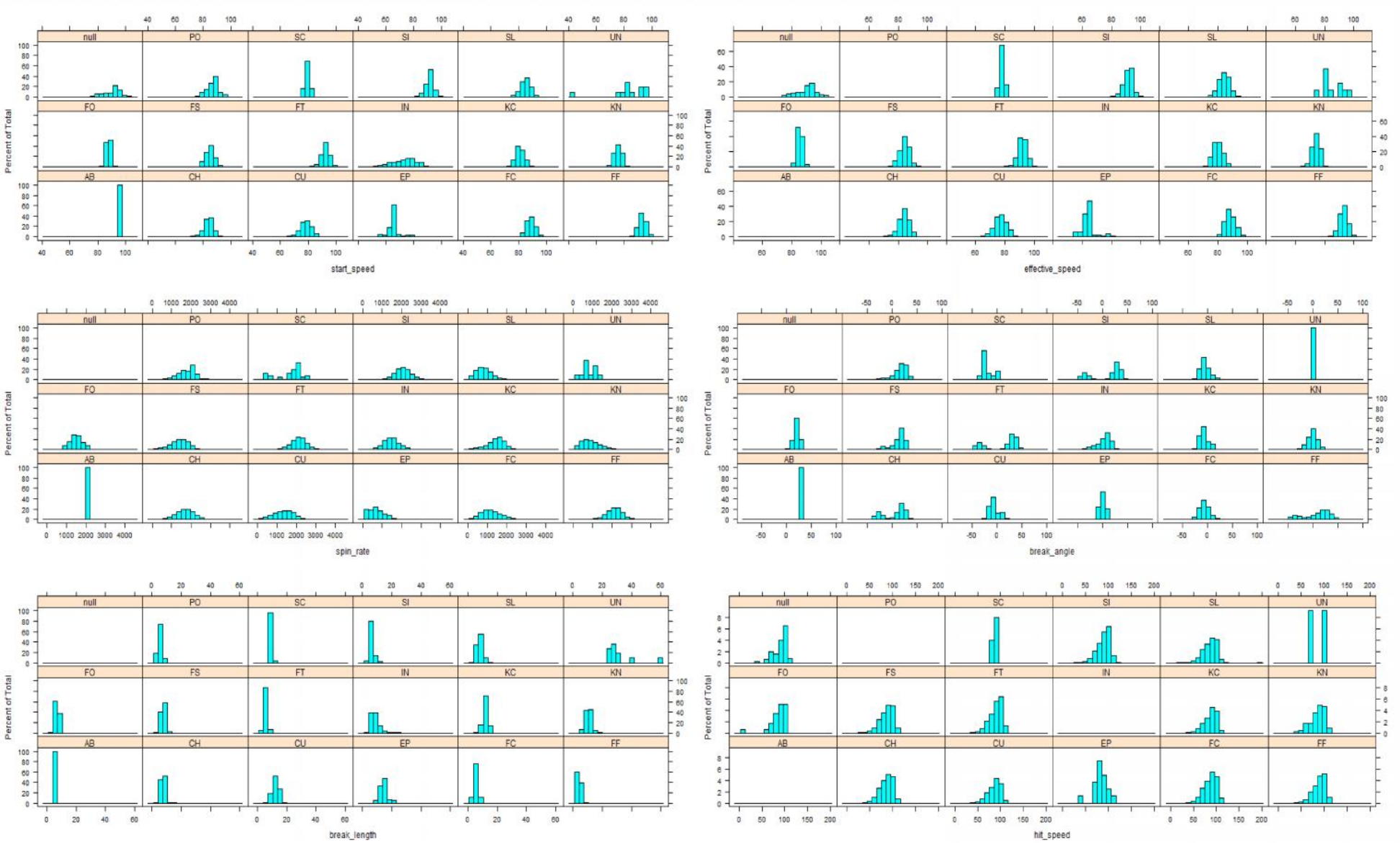
Figure 2: Histograms



The distributions deviate from normality even further when plotted by pitch type. These plots are displayed in Figure 3 on the following page. The histograms for hit speed, located in the bottom right corner of Figure 3, vary considerably between pitch types. With hit speed being the dependent variable of interest, the normality assumption was violated and the ANOVA test was ruled out as a means to investigate the differences between groups. The Kruskal-Wallis Test, a rank-based nonparametric test, was utilized instead.

Unlike the ANOVA test, the Kruskal-Wallis Test is not based on any assumption about the shape of the distribution. Therefore, the non-normality displayed in the histograms is acceptable. One assumption that must be met is that the dependent variable should be measured at the ordinal or continuous level. In this case, the dependent variable was measured at the continuous (i.e. ratio) level. In addition, this test is commonly used when there are 3 or more groups to compare. In this case, there were 14 different pitch types present in the dataset. Finally, observations in each group or between the groups themselves must be independent. Since each observation in the dataset represented a unique pitch that occurred during the 2016 MLB season, this assumption was also met.

Figure 3: Histograms by pitch type



Results

The Kruskal-Wallis Test for hit speed by pitch type produced a chi-squared test statistic of 1118.9 and p-value less than 2.2e-16. Compared to the critical chi-square value for a 95% confidence interval with 14 degrees of freedom of 23.7, this test statistic provided sufficient evidence to reject the null hypothesis that all groups are identical. Therefore, it was concluded that the hit speed for at least one group differs from the others. The mean hit speeds for each pitch type are shown below in Table 3.

Table 3: Mean hit speed by pitch type

pitch_type	EP	SC	CH	KN	SL	UN	CU	FO	FC	FS	KC	FF	SI	FT
hit_speed	82.25	85.08	86.06	86.18	86.27	86.32	86.34	86.35	86.71	86.88	87.76	89.21	89.53	89.79

However, this test does not indicate in which groups the difference lies. In order to determine which specific pitch types presented a statistically significant different in hit speed, a post-hoc pairwise comparison test was performed on the data. Selected results are presented below in Table 4, with the full results included in Appendix B.

Table 4: Selected results of Kruskal-Wallis pairwise comparisons for hit speed by pitch type

95% Confidence intervals for Kruskal-Wallis comparisons						
Pair	Diff	Lower	Upper	Adj. P-value	Decision	Interpretation
CH & FF	-5842.13	-6923.27	-4761.00	0	Reject H0	CH mean hit speed is lower than FF
CU & FF	-5424.70	-6725.64	-4123.75	0	Reject H0	CU mean hit speed is lower than FF
FC & FF	-4639.09	-6072.79	-3205.39	0	Reject H0	FC mean hit speed is lower than FF
FF & FS	4007.57	1412.81	6602.33	0.000004	Reject H0	FS mean hit speed is lower than FF
CH & FT	-7244.67	-8475.95	-6013.38	0	Reject H0	CH mean hit speed is lower than FT
CU & FT	-6827.23	-8255.40	-5399.06	0	Reject H0	CU mean hit speed is lower than FT
FC & FT	-6041.63	-7591.70	-4491.55	0	Reject H0	FC mean hit speed is lower than FT
FF & FT	-1402.53	-2354.35	-450.72	0.000018	Reject H0	FF mean hit speed is lower than FT
FS & FT	-5410.10	-8070.93	-2749.27	0	Reject H0	FS mean hit speed is lower than FT
CH & KC	-2858.95	-5381.18	-336.73	0.007108	Reject H0	CH mean hit speed is lower than KC
FF & KC	2983.18	584.94	5381.41	0.0012	Reject H0	KC mean hit speed is lower than FF
FT & KC	4385.71	1916.15	6855.27	0	Reject H0	KC mean hit speed is lower than FT
FF & KN	4665.66	988.52	8342.79	0.000786	Reject H0	KN mean hit speed is lower than FF
FT & KN	6068.19	2344.14	9792.24	0.000001	Reject H0	KN mean hit speed is lower than FT
CH & SI	-6832.86	-8317.94	-5347.77	0	Reject H0	CH mean hit speed is lower than SI
CU & SI	-6415.42	-8067.42	-4763.43	0	Reject H0	CU mean hit speed is lower than SI
FC & SI	-5629.82	-7388.26	-3871.37	0	Reject H0	FC mean hit speed is lower than SI
FS & SI	-4998.29	-7785.66	-2210.93	0	Reject H0	FS mean hit speed is lower than SI
KC & SI	-3973.90	-6579.31	-1368.50	0.000006	Reject H0	KC mean hit speed is lower than SI
KN & SI	-5656.38	-9471.87	-1840.90	0.000015	Reject H0	KN mean hit speed is lower than SI
FF & SL	5375.99	4390.89	6361.10	0	Reject H0	SL mean hit speed is lower than FF
FT & SL	6778.53	5630.64	7926.42	0	Reject H0	SL mean hit speed is lower than FT
SI & SL	6366.72	4950.01	7783.43	0	Reject H0	SL mean hit speed is lower than SI

Table 4 reveals 23 out of 91 pairwise comparisons that produced a p-value sufficient to reject the null hypothesis and conclude that there was a statistically significant difference in hit speed between the

two pitch types. The interpretation was derived by comparing the mean hit speeds shown in Table 3 for the two pitch types in question. For example, mean hit speeds for the changeup (CH) and four-seam fastball (FF) were 86.06 and 89.21, respectively. Therefore, it was concluded that the changeup generated a lower hit speed than the four-seam fastball. Most of the pairs shown above represent a comparison between a slower off-speed pitch like the changeup (CH), knuckleball (KN), or slider (SL) and faster pitch like the four-seam fastball (FF), sinker (SI), or two-seam fastball (FT).

Hit speed indicates only how hard the ball was hit by the batter with no indication of the result of the play. In order to assess performance from an outcome perspective, the statistic Batting Average on Balls In Play (BABIP) was calculated for each pitch type. Using the "event" variable within the dataset that indicates the result of each at bat, a subset was taken to include only observations with a hit speed greater than zero and only one observation per at bat (i.e. one record for every at bat that ended with a ball put in play). The aggregate batting average for each pitch type was calculated as total number of hits divided by total number of at bats. The results are displayed below in Table 5.

Table 5: BABIP by pitch type

pitch_type	EP	KN	CH	SL	FS	FC	CU	FT	KC	SI	FF	FO	UN	SC
BABIP	0.294	0.311	0.342	0.344	0.345	0.346	0.349	0.363	0.366	0.366	0.372	0.375	0.500	0.667

This value was combined with the batted ball dataset such that each observation reflected the BABIP for its associated pitch type and a Kruskal-Wallis Test was performed. The test produced a chi-squared test statistic of 84620 and p-value less than 2.2e-16. Therefore, it was concluded that the BABIP for at least one pitch type differs from the others. The same post-hoc pairwise comparison was completed and selected results are shown below in Table 6, with the full results included in Appendix B.

Table 6: Selected results of Kruskal-Wallis pairwise comparisons for BABIP by pitch type

95% Confidence intervals for Kruskal-Wallis comparisons						
Pair	Diff	Lower	Upper	Adj. P-value	Decision	Interpretation
CH & CU	-25717.50	-27132.30	-24302.70	0	Reject H0	CH BABIP is lower than CU
CH & FT	-35418.00	-36566.66	-34269.34	0	Reject H0	CH BABIP is lower than FT
CU & FT	-9700.50	-11031.37	-8369.63	0	Reject H0	CU BABIP is lower than FT
FC & FT	-15013.50	-16467.39	-13559.61	0	Reject H0	FC BABIP is lower than FT
FF & FT	29182.50	28290.40	30074.60	0	Reject H0	FT BABIP is lower than FF
CU & KC	-20374.50	-22802.59	-17946.41	0	Reject H0	CU BABIP is lower than KC
CH & KN	5110.00	1566.52	8653.48	0.000034	Reject H0	KN BABIP is lower than CH
CU & SI	-20374.50	-21917.22	-18831.78	0	Reject H0	CU BABIP is lower than SI
FF & SI	18508.50	17323.34	19693.66	0	Reject H0	SI BABIP is lower than FF
FT & SI	-10674.00	-11976.97	-9371.03	0	Reject H0	FT BABIP is lower than SI
CH & SL	-10760.00	-11935.00	-9585.00	0	Reject H0	CH BABIP is lower than SL
CU & SL	14957.50	13603.83	16311.17	0	Reject H0	SL BABIP is lower than CU
FC & SL	9644.50	8169.71	11119.29	0	Reject H0	SL BABIP is lower than FC
KC & SL	35332.00	33035.35	37628.65	0	Reject H0	SL BABIP is lower than KC
SI & SL	35332.00	34005.75	36658.25	0	Reject H0	SL BABIP is lower than SI

In this case, 73 out of the 91 possible pairs produced a p-value low enough to reject the null hypothesis and conclude that there was a statistically significant difference in BABIP between the two pitch types. Again, the interpretation was derived by comparing the BABIPs shown in Table 5 for the two pitch types in question. One noteworthy comparison was between the changeup (CH) and the slider (SL). It was revealed that BABIP for the changeup (CH) was significantly different than BABIP for the slider (SL) despite a difference of only 0.002 in those values. Likewise, the sinker (SI) showed a statistically significant difference from the four-seam fastball (FF). Overall, slower off-speed pitches like the changeup (CH) or slider (SL) produced BABIPs lower than faster pitches like the sinker (SI) or four-seam fastball (FF). Nearly all of the 18 comparisons that failed to reject the null hypothesis included the eephus (EP), the screwball (SC), or unknown (UN) pitch types. These pitch types are either rarely used "trick" pitches or failures by the PITCHf/x system to classify the pitch. Therefore, the practical interpretation of this test is that while the differences in BABIP between each commonly used pitch type may be small, they are likely to be statistically significant.

Implications

The hypothesis that slower pitch types such as the curveball or changeup would perform better on batted ball hit speed proved to be true. The curveball (CU), changeup (CH), and slider (SL) each displayed a significantly lower hit speed than the four-seam fastball (FF), two-seam fastball (FT), and sinker (SI). In other words, a pitch thrown at a slower speed resulted in the ball being struck by the batter at a slower speed. The hypothesis that the sinker would perform better on BABIP proved to be false. Rather, it produced a higher BABIP than most pitches, including all off-speed pitch types, and pairwise comparisons revealed these differences to be statistically significant. Based on the results of this study, slower pitch types resulted in fewer hits than faster ones.

These results could impact areas such as scouting, roster construction, and player development. Teams may seek out players who specialize in throwing a curveball, changeup, or slider rather than players who throw a high-speed fastball. They may also try to teach these pitch types to their prospects in the minor leagues to develop more effective pitchers for the major league team. In-game strategy may also be impacted. Pitchers may choose to throw a curveball or changeup rather than a fastball, knowing that these pitch types induce weaker contact and a lower batting average.

One limitation of this study is that it did not include historical data. The PITCHf/x database contains a massive amount of data and analyzing all of it requires more computing resources than were available. It would be interesting to investigate how the performance of different pitch types has changed over time. The recent preference across MLB has been for pitchers who throw faster, so one possible explanation for the results of this study is that batters have adjusted to faster pitches and are less prepared to hit slower pitches. Also, it did not distinguish between left-handed and right-handed throwers. Pitch types behave differently (i.e. break in different directions) depending on which hand the pitcher throws with, so this may have changed the results.

Appendix A: References

Baseball Prospectus | Glossary. (2016). Retrieved July 10, 2016, from http://www.baseballprospectus.com/glossary/

Berry, A. (2013, April 26). Hitters and pitchers agree: There is no pitch tougher to hit in baseball than a well placed fastball. Retrieved August 26, 2016, from http://m.mlb.com/news/article/45834916/hitters-and-pitchers-agree-there-is-no-pitch-tougher-to-hit-in-baseball-than-a-well-placed-fastball/

Digging Into the Data Behind Baseball's 10 Toughest Pitches. (2015). Retrieved August 26, 2016, from http://www.stats.com/insights/mlb/digging-into-the-data-behind-baseballs-10-toughest-pitches/

Fast, M. (2007, August 02). Glossary of the Gameday pitch fields. Retrieved July 10, 2016, from https://fastballs.wordpress.com/2007/08/02/glossary-of-the-gameday-pitch-fields/

Fast, M. (2010, April 5). Lose a tick, gain a tick. Retrieved August 26, 2016, from http://www.hardballtimes.com/lose-a-tick-gain-a-tick/

Glossary. (2016). Retrieved July 10, 2016, from http://m.mlb.com/glossary/statcast/

Marchi, M., & Albert, J. (2013). Analyzing baseball data with R. Boca Raton, FL: CRC Press.

Pitches - BR Bullpen. (2016, May 5). Retrieved August 26, 2016, from http://www.baseball-reference.com/bullpen/pitches

Sarris, E. (2016, January 25). Last Year's Best Pitch By the Numbers | FanGraphs | FanGraphs Baseball. Retrieved August 26, 2016, from http://www.fangraphs.com/plus/last-years-best-pitch-by-the-numbers/

Sullivan, J. (2013, November 11). Identifying 2013's Most Unhittable Pitches. Retrieved August 26, 2016, from http://www.fangraphs.com/blogs/identifying-2013s-most-unhittable-pitches/

Appendix B: Expanded Tables and Figures

Pitch Type Glossary

Pitch Type	PITCHf/x Abbreviation	Description
Automatic Ball	AB	Automatic ball called by the umpire due to a rules violation by the pitcher
Changeup	CH	A slow pitch thrown with the same arm motion as the fastball but with less velocity
Curveball	CU	A slower pitch thrown with spin that creates a straight downward break
Eephus	EP	An incredibly slow pitch with a very high arching trajectory
Fastball (unspecified)	FA	A fastball that PITCHf/x was not able to classify further into four-seam, two-seam, or cut
Cut-fastball (cutter)	FC	A variation of the fastball that is similar to the slider, but with more velocity and less movement
Four-seam Fastball	FF	The standard pitch in baseball, typically thrown the fastest with little or no movement
Forkball	FO	A slower offspeed pitch that tumbles - rather than breaks - downward
Splitter	FS	Similar to a sinker, meant to look like a fastball then dive downward
Two-seam Fastball	FT	A variation of the fastball that has slightly less speed but more downward and/or tailing movement
Intentional Ball	IN	A pitch thrown well outside the strike zone with the intention of walking the batter
Knuckle-curve	KC	A variation of the curveball that is slightly slower
Knuckleball	KN	A trick pitch thrown with no spin that tumbles in various directions
Pitchout	PO	A pitch thrown well outside the strike zone with the intention of catching a baserunner attempting to steal
Screwball	SC	Similar in trajectory to a curveball, but breaks the opposite direction - known as the "backwards curveball"
Sinker	SI	The same thing as the two-seam fastball - a variation of the fastball with less speed but more downward and/or tailing movement
Slider	SL	A breaking ball thrown faster than the curveball with less vertical and more horizontal break
Unknown	UN	PITCHf/x failed to capture data or was unable to identify the pitch
		Sources: (Marchi, 2013), (Pitches - BR Bullpen, 2015)

Table 7: Full results of Kruskal-Wallis pairwise comparisons for hit speed by pitch type

95% Confidence intervals for Kruskal-Wallis comparisons					
Pair	**Diff**	**Lower**	**Upper**	**Adj. P-value**	**Decision**
CH & CU	-417.43	-1934.86	1099.99	1	FTR H0
CH & EP	7125.80	-15806.58	30058.18	1	FTR H0
CU & EP	7543.23	-15400.55	30487.02	1	FTR H0
CH & FC	-1203.04	-2835.71	429.63	1	FTR H0
CU & FC	-785.61	-2571.45	1000.23	1	FTR H0
EP & FC	-8328.84	-31280.54	14622.86	1	FTR H0
CH & FF	-5842.13	-6923.27	-4761.00	0	Reject H0

CU & FF	-5424.70	-6725.64	-4123.75	0	Reject H0
EP & FF	-12967.93	-35887.00	9951.14	1	FTR H0
FC & FF	-4639.09	-6072.79	-3205.39	0	Reject H0
CH & FO	-3583.85	-21152.31	13984.60	1	FTR H0
CU & FO	-3166.42	-20749.77	14416.93	1	FTR H0
EP & FO	-10709.65	-39567.33	18148.02	1	FTR H0
FC & FO	-2380.81	-19974.49	15212.86	1	FTR H0
FF & FO	2258.28	-15292.81	19809.36	1	FTR H0
CH & FS	-1834.56	-4544.34	875.21	1	FTR H0
CU & FS	-1417.13	-4221.86	1387.60	1	FTR H0
EP & FS	-8960.37	-32013.73	14093.00	1	FTR H0
FC & FS	-631.52	-3500.24	2237.20	1	FTR H0
FF & FS	4007.57	1412.81	6602.33	0.000004	Reject H0
FO & FS	1749.29	-15976.81	19475.39	1	FTR H0
CH & FT	-7244.67	-8475.95	-6013.38	0	Reject H0
CU & FT	-6827.23	-8255.40	-5399.06	0	Reject H0
EP & FT	-14370.47	-37297.11	8556.18	1	FTR H0
FC & FT	-6041.63	-7591.70	-4491.55	0	Reject H0
FF & FT	-1402.53	-2354.35	-450.72	0.000018	Reject H0
FO & FT	-3660.81	-21221.79	13900.16	1	FTR H0
FS & FT	-5410.10	-8070.93	-2749.27	0	Reject H0
CH & KC	-2858.95	-5381.18	-336.73	0.007108	Reject H0
CU & KC	-2441.52	-5065.49	182.45	0.126694	FTR H0
EP & KC	-9984.76	-33016.83	13047.32	1	FTR H0
FC & KC	-1655.91	-4348.18	1036.35	1	FTR H0
FF & KC	2983.18	584.94	5381.41	0.0012	Reject H0
FO & KC	724.90	-16973.50	18423.30	1	FTR H0
FS & KC	-1024.39	-4477.73	2428.95	1	FTR H0
FT & KC	4385.71	1916.15	6855.27	0	Reject H0
CH & KN	-1176.47	-4935.65	2582.71	1	FTR H0
CU & KN	-759.04	-4587.23	3069.15	1	FTR H0
EP & KN	-8302.27	-31502.41	14897.86	1	FTR H0
FC & KN	26.57	-3848.75	3901.88	1	FTR H0
FF & KN	4665.66	988.52	8342.79	0.000786	Reject H0
FO & KN	2407.38	-15509.18	20323.94	1	FTR H0
FS & KN	658.09	-3779.86	5096.04	1	FTR H0
FT & KN	6068.19	2344.14	9792.24	0.000001	Reject H0
KC & KN	1682.48	-2643.50	6008.46	1	FTR H0
CH & SC	6929.43	-47622.51	61481.37	1	FTR H0
CU & SC	7346.86	-47209.88	61903.60	1	FTR H0
EP & SC	-196.37	-59357.42	58964.67	1	FTR H0
FC & SC	8132.47	-46427.60	62692.54	1	FTR H0
FF & SC	12771.56	-41774.79	67317.91	1	FTR H0
FO & SC	10513.28	-46782.32	67808.88	1	FTR H0
FS & SC	8763.99	-45838.92	63366.91	1	FTR H0

FT & SC	14174.09	-40375.44	68723.63	1	FTR H0
KC & SC	9788.38	-44805.55	64382.31	1	FTR H0
KN & SC	8105.90	-46559.14	62770.94	1	FTR H0
CH & SI	-6832.86	-8317.94	-5347.77	0	Reject H0
CU & SI	-6415.42	-8067.42	-4763.43	0	Reject H0
EP & SI	-13958.66	-36900.33	8983.01	1	FTR H0
FC & SI	-5629.82	-7388.26	-3871.37	0	Reject H0
FF & SI	-990.73	-2253.80	272.35	0.738619	FTR H0
FO & SI	-3249.00	-20829.60	14331.59	1	FTR H0
FS & SI	-4998.29	-7785.66	-2210.93	0	Reject H0
FT & SI	411.81	-981.96	1805.57	1	FTR H0
KC & SI	-3973.90	-6579.31	-1368.50	0.000006	Reject H0
KN & SI	-5656.38	-9471.87	-1840.90	0.000015	Reject H0
SC & SI	-13762.29	-68318.14	40793.56	1	FTR H0
CH & SL	-466.14	-1723.34	791.06	1	FTR H0
CU & SL	-48.70	-1499.27	1401.86	1	FTR H0
EP & SL	-7591.94	-30519.99	15336.11	1	FTR H0
FC & SL	736.90	-833.83	2307.63	1	FTR H0
FF & SL	5375.99	4390.89	6361.10	0	Reject H0
FO & SL	3117.71	-14445.09	20680.52	1	FTR H0
FS & SL	1368.43	-1304.49	4041.34	1	FTR H0
FT & SL	6778.53	5630.64	7926.42	0	Reject H0
KC & SL	2392.82	-89.77	4875.40	0.081683	FTR H0
KN & SL	710.34	-3022.36	4443.03	1	FTR H0
SC & SL	-7395.57	-61945.69	47154.56	1	FTR H0
SI & SL	6366.72	4950.01	7783.43	0	Reject H0
CH & UN	-4550.16	-71359.04	62258.73	1	FTR H0
CU & UN	-4132.72	-70945.52	62680.08	1	FTR H0
EP & UN	-11675.96	-82298.49	58946.57	1	FTR H0
FC & UN	-3347.11	-70162.63	63468.40	1	FTR H0
FF & UN	1291.98	-65512.34	68096.29	1	FTR H0
FO & UN	-966.30	-70033.65	68101.05	1	FTR H0
FS & UN	-2715.59	-69566.10	64134.92	1	FTR H0
FT & UN	2694.51	-64112.40	69501.43	1	FTR H0
KC & UN	-1691.20	-68534.37	65151.97	1	FTR H0
KN & UN	-3373.68	-70274.95	63527.58	1	FTR H0
SC & UN	-11479.58	-97720.88	74761.72	1	FTR H0
SI & UN	2282.70	-64529.37	69094.78	1	FTR H0
SL & UN	-4084.02	-70891.41	62723.38	1	FTR H0

Table 8: Full results of Kruskal-Wallis pairwise comparisons for BABIP by pitch type

95% Confidence intervals for Kruskal-Wallis comparisons					
Pair	**Diff**	**Lower**	**Upper**	**Adj. P-value**	**Decision**
CH & CU	-25717.50	-27132.30	-24302.70	0	Reject H0

CH & EP	5429.00	-15334.89	26192.89	1	FTR H0
CU & EP	31146.50	10371.74	51921.26	0.000011	Reject H0
CH & FC	-20404.50	-21935.60	-18873.40	0	Reject H0
CU & FC	5313.00	3640.86	6985.14	0	Reject H0
EP & FC	-25833.50	-46616.51	-5050.49	0.001218	Reject H0
CH & FF	-64600.50	-65613.57	-63587.43	0	Reject H0
CU & FF	-38883.00	-40098.77	-37667.23	0	Reject H0
EP & FF	-70029.50	-90780.78	-49278.22	0	Reject H0
FC & FF	-44196.00	-45545.33	-42846.67	0	Reject H0
CH & FO	-79166.00	-96295.58	-62036.42	0	Reject H0
CU & FO	-53448.50	-70591.26	-36305.74	0	Reject H0
EP & FO	-84595.00	-111483.88	-57706.12	0	Reject H0
FC & FO	-58761.50	-75914.25	-41608.75	0	Reject H0
FF & FO	-14565.50	-31679.79	2548.79	0.340951	FTR H0
CH & FS	-17436.50	-19989.47	-14883.53	0	Reject H0
CU & FS	8281.00	5641.03	10920.97	0	Reject H0
EP & FS	-22865.50	-43748.67	-1982.33	0.012774	Reject H0
FC & FS	2968.00	263.93	5672.07	0.012272	Reject H0
FF & FS	47164.00	44715.72	49612.28	0	Reject H0
FO & FS	61729.50	44455.52	79003.48	0	Reject H0
CH & FT	-35418.00	-36566.66	-34269.34	0	Reject H0
CU & FT	-9700.50	-11031.37	-8369.63	0	Reject H0
EP & FT	-40847.00	-61605.34	-20088.66	0	Reject H0
FC & FT	-15013.50	-16467.39	-13559.61	0	Reject H0
FF & FT	29182.50	28290.40	30074.60	0	Reject H0
FO & FT	43748.00	26625.15	60870.85	0	Reject H0
FS & FT	-17981.50	-20488.92	-15474.08	0	Reject H0
CH & KC	-46092.00	-48425.21	-43758.79	0	Reject H0
CU & KC	-20374.50	-22802.59	-17946.41	0	Reject H0
EP & KC	-51521.00	-72378.45	-30663.55	0	Reject H0
FC & KC	-25687.50	-28185.15	-23189.85	0	Reject H0
FF & KC	18508.50	16290.33	20726.67	0	Reject H0
FO & KC	33074.00	15831.13	50316.87	0	Reject H0
FS & KC	-28655.50	-31882.20	-25428.80	0	Reject H0
FT & KC	-10674.00	-12957.29	-8390.71	0	Reject H0
CH & KN	5110.00	1566.52	8653.48	0.000034	Reject H0
CU & KN	30827.50	27220.84	34434.16	0	Reject H0
EP & KN	-319.00	-21346.25	20708.25	1	FTR H0
FC & KN	25514.50	21860.65	29168.35	0	Reject H0
FF & KN	69710.50	66241.69	73179.31	0	Reject H0
FO & KN	84276.00	66828.11	101723.89	0	Reject H0
FS & KN	22546.50	18360.33	26732.67	0	Reject H0
FT & KN	40528.00	37017.19	44038.81	0	Reject H0
KC & KN	51202.00	47146.11	55257.89	0	Reject H0
6:46 PM	-79182.00	-128573.35	-29790.65	0.000001	Reject H0

CU & SC	-53464.50	-102860.42	-4068.58	0.015399	Reject H0
EP & SC	-84611.00	-138174.93	-31047.07	0.000002	Reject H0
FC & SC	-58777.50	-108176.89	-9378.11	0.002922	Reject H0
FF & SC	-14581.50	-63967.54	34804.54	1	FTR H0
FO & SC	-16.00	-52278.58	52246.58	1	FTR H0
FS & SC	-61745.50	-111187.11	-12303.89	0.001104	Reject H0
FT & SC	-43764.00	-93153.01	5625.01	0.222353	FTR H0
KC & SC	-33090.00	-82520.75	16340.75	1	FTR H0
KN & SC	-84292.00	-133794.64	-34789.36	0	Reject H0
CH & SI	-46092.00	-47480.59	-44703.41	0	Reject H0
CU & SI	-20374.50	-21917.22	-18831.78	0	Reject H0
EP & SI	-51521.00	-72294.00	-30748.00	0	Reject H0
FC & SI	-25687.50	-27337.52	-24037.48	0	Reject H0
FF & SI	18508.50	17323.34	19693.66	0	Reject H0
FO & SI	33074.00	15933.38	50214.62	0	Reject H0
FS & SI	-28655.50	-31281.51	-26029.49	0	Reject H0
FT & SI	-10674.00	-11976.97	-9371.03	0	Reject H0
KC & SI	0.00	-2412.92	2412.92	1	FTR H0
KN & SI	-51202.00	-54798.46	-47605.54	0	Reject H0
SC & SI	33090.00	-16305.17	82485.17	1	FTR H0
CH & SL	-10760.00	-11935.00	-9585.00	0	Reject H0
CU & SL	14957.50	13603.83	16311.17	0	Reject H0
EP & SL	-16189.00	-36948.81	4570.81	0.777119	FTR H0
FC & SL	9644.50	8169.71	11119.29	0	Reject H0
FF & SL	53840.50	52914.73	54766.27	0	Reject H0
FO & SL	68406.00	51281.36	85530.64	0	Reject H0
FS & SL	6676.50	4156.90	9196.10	0	Reject H0
FT & SL	24658.00	23585.54	25730.46	0	Reject H0
KC & SL	35332.00	33035.35	37628.65	0	Reject H0
KN & SL	-15870.00	-19389.52	-12350.48	0	Reject H0
SC & SL	68422.00	19032.37	117811.63	0.000098	Reject H0
SI & SL	35332.00	34005.75	36658.25	0	Reject H0
CH & UN	-79179.50	-139668.09	-18690.91	0.000392	Reject H0
CU & UN	-53462.00	-113954.33	7030.33	0.22886	FTR H0
EP & UN	-84608.50	-148549.56	-20667.44	0.000304	Reject H0
FC & UN	-58775.00	-119270.16	1720.16	0.073651	FTR H0
FF & UN	-14579.00	-75063.27	45905.27	1	FTR H0
FO & UN	-13.50	-62868.43	62841.43	1	FTR H0
FS & UN	-61743.00	-122272.64	-1213.36	0.037828	Reject H0
FT & UN	-43761.50	-104248.19	16725.19	1	FTR H0
KC & UN	-33087.50	-93608.27	27433.27	1	FTR H0
KN & UN	-84289.50	-144869.00	-23710.00	0.000088	Reject H0
SC & UN	2.50	-78079.67	78084.67	1	FTR H0
SI & UN	-33087.50	-93579.22	27404.22	1	FTR H0
SL & UN	-68419.50	-128906.70	-7932.30	0.007372	Reject H0

Appendix C: R Code

```
# PREDICT 456 Sports Performance Analysis Section 55 Summer 2016
# Christopher Anderson
# Assignment #4
# Last modified August 28, 2016

library(moments)
library(ggplot2)
library(gridExtra)
library(pitchRx)
library(Hmisc)
library(psych)
library(lattice)
library(FSA)
library(asbio)
library(dplyr)

# Read in PITCHf/x data downloaded from MLB Statcast Search tool
mydata <- read.csv("savant_data.csv", header = T, sep = ",")

# Subset the data to include only relevant columns (541581 obs. of 12 variables)
mydata <- subset(mydata,
select=c("pitch_type","description","events","type","start_speed","effective_speed",

"spin_rate","break_angle","break_length","hit_speed","stand","p_throws"))

# Remove records with missing data
mydata <- na.omit(mydata) # 528486 obs. of  12 variables; 13095 obs. removed

# Examine structure of data
str(mydata)

# Amend data types
mydata$start_speed <- as.numeric(as.character(mydata$start_speed))
mydata$effective_speed <- as.numeric(as.character(mydata$effective_speed))
mydata$spin_rate <- as.numeric(as.character(mydata$spin_rate))
mydata$break_angle <- as.numeric(as.character(mydata$break_angle))
mydata$break_length <- as.numeric(as.character(mydata$break_length))
mydata$hit_speed <- as.numeric(as.character(mydata$hit_speed))
str(mydata)
head(mydata)
tail(mydata)

# Summary statistics by pitch type
aggregate(cbind(start_speed, effective_speed, spin_rate, break_angle, break_length,
hit_speed) ~ pitch_type, data = mydata, mean)
summary(mydata$pitch_type)

# Exploratory boxplots of pitch measurements by pitch type
plot1 <- ggplot(data = mydata, aes(x = pitch_type, y = start_speed, group = pitch_type,
                                   colour = pitch_type))+ geom_boxplot(show.legend = FALSE)
plot2 <- ggplot(data = mydata, aes(x = pitch_type, y = effective_speed, group = pitch_type,
```

```
                                    colour = pitch_type))+ geom_boxplot(show.legend = FALSE)
plot3 <- ggplot(data = mydata, aes(x = pitch_type, y = spin_rate, group = pitch_type,
                                    colour = pitch_type))+ geom_boxplot(show.legend = FALSE)
plot4 <- ggplot(data = mydata, aes(x = pitch_type, y = break_angle, group = pitch_type,
                                    colour = pitch_type))+ geom_boxplot(show.legend = FALSE)
plot5 <- ggplot(data = mydata, aes(x = pitch_type, y = break_length, group = pitch_type,
                                    colour = pitch_type))+ geom_boxplot(show.legend = FALSE)
plot6 <- ggplot(data = mydata, aes(x = pitch_type, y = hit_speed, group = pitch_type,
                                    colour = pitch_type))+ geom_boxplot(show.legend = FALSE)
grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, ncol=2)

# Exploratory histograms and normality assessment
par(mfrow = c(3,2), oma=c(0,0,2,0))
hist(mydata$start_speed, xlab = "Start Speed", main = "", col = "steelblue")
hist(mydata$effective_speed, xlab = "Effective Speed", main = "", col = "steelblue")
hist(mydata$spin_rate, xlab = "Spin Rate", main = "", col = "steelblue")
hist(mydata$break_angle, xlab = "Break Angle", main = "", col = "steelblue")
hist(mydata$break_length, xlab = "Break Length", main = "", col = "steelblue")
hist(mydata$hit_speed, xlab = "Hit Speed", main = "", col = "steelblue")
title("", outer=TRUE)
par(mfrow = c(1,1))

# Exploratory histograms within each pitch type
plot1 <- histogram(~ start_speed | pitch_type, data=mydata)
plot2 <- histogram(~ effective_speed | pitch_type, data=mydata)
plot3 <- histogram(~ spin_rate | pitch_type, data=mydata)
plot4 <- histogram(~ break_angle | pitch_type, data=mydata)
plot5 <- histogram(~ break_length | pitch_type, data=mydata)
plot6 <- histogram(~ hit_speed | pitch_type, data=mydata)
grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, ncol=2)

# Kruskal-Wallis test for hit speed by pitch type
qchisq(0.950, 14) # Critical chi-squared value = 23.68479
kruskal.test(hit_speed ~ pitch_type, data = mydata) # Kruskal-Wallis chi-squared = 1118.9,
df = 14, p-value < 2.2e-16
pairw.kw(y = mydata$hit_speed, x = mydata$pitch_type, conf = .95)

# Batting average on balls in play
# Filter to include only one record per at bat
BABIPdata <-
  filter(mydata, description == "In play, no out"| description == "In play, out(s)"|
description == "In play, run(s)")
  # 95723 obs. of 12 variables
BABIPdata <- filter(BABIPdata, hit_speed > 0) # 84621 obs. of 12 variables; 11102 obs.
removed

# Identify each observation as a Hit or not
H <- c("Single", "Double", "Triple", "Home Run")

# Identify each observationl as an official at bat or not (events like walks and hit by
pitches are excluded)
AB <- c("Single", "Double", "Triple", "Home Run","Bunt Groundout","Bunt Lineout","Bunt Pop
Out","Double Play","Field Error",
```

```
        "Fielders Choice","Fielders Choice Out","Flyout","Forceout","Grounded Into
DP","Groundout","Lineout","Pop Out",
        "Strikeout","Strikeout - DP","Triple Play")

# Calculate BABIP per pitch type and put results in dataframe
Hit <- factor(BABIPdata$events %in% H, labels = c(0, 1))
AtBats <- factor(BABIPdata$events %in% AB, labels = c(0, 1))
BABIPdata <- data.frame(BABIPdata, Hit, AtBats) # 84621 obs. of 14 variables

BABIPdata$Hit <- as.numeric(as.character(BABIPdata$Hit))
BABIPdata$AtBats <- as.numeric(as.character(BABIPdata$AtBats))

Hit.type <- aggregate(Hit ~ pitch_type, data = BABIPdata, sum)
AtBats.type <- aggregate(AtBats ~ pitch_type, data = BABIPdata, sum)
BABIP <- round(Hit.type$Hit / AtBats.type$AtBats, 3)
BABIP <- data.frame(Hit.type$pitch_type, BABIP)
colnames(BABIP) <- c("pitch_type","BABIP")
BABIPdata <- merge(BABIPdata, BABIP, by = 'pitch_type')

# View BABIP by pitch type
BABIP[order(BABIP$BABIP, decreasing = FALSE),]

# Kruskal-Wallis test for BABIP by pitch type
kruskal.test(BABIP ~ pitch_type, data = BABIPdata)
pairw.kw(y = BABIPdata$BABIP, x = BABIPdata$pitch_type, conf = .95)


# Save datasets for future use
write.csv(mydata, file = "assignment4_mydata.csv")
write.csv(BABIPdata, file = "assignment4_BABIPdata.csv")

# End
```